Curriculum vitae e attività di ricerca di Matteo Rè

Aggiornato al 2 dicembre 2015

e-mail: re@di.unimi.it

home page: http://homes.di.unimi.it/re/

Matteo Rè si è laureato in Biologia (indirizzo di specializzazione: biologia molecolare, sottoarea di indirizzo: bioinformatica) all'Università di Milano nel 2004. Nel 2007 ha conseguito il dottorato in Biologia Cellulare e Molecolare all'Università di Milano, discutendo una tesi dal titolo "Development of machine learning methods for the discrimination between coding and non-coding conserved sequences".

Nel 2005 ha ottenuto una borsa di ricerca dall' Associazione Italiana per la Ricerca sul Cancro (AIRC). Nel 2008 ha collaborato come ricercatore a contratto presso l'unità di ricerca di genetica delle malattie comuni presso il dipartimento di biotecnologie della fondazione S.Raffaele di Milano, svolgendo attività di ricerca in bioinformatica. Dal 01/09/2008 al 31/10/2009 e dal 01/11/2009 al 31/10/2011 è stato assegnista presso il DSI, Dipartimento di Scienze dell'Informazione, dell'Università degli Studi di Milano. Durante il 2012 è stato ricercatore a contratto presso la Fondazione "Filarete" di Milano per l'elaborazione computazionale di dati next-generation sequencing. Dal Novembre 2012 al mese di Ottobre 2015 è stato ricercatore a tempo determinato (di tipo A) presso il dipartimento di Informatica dell'Università degli Studi di Milano come membro del gruppo di ricerca in biologia computazionale coordinato dal prof. Giorgio Valentini ed è stato docente in numerosi corsi di laurea triennale e magistrale nei corsi di laurea in Informatica, Biotecnologie Industriali e Ambientali, Biotecnologie del Farmaco. Ha seguito come relatore e co-relatore tesi di laurea magistrale in Informatica, Matematica applicata e Biotecnologie Industriali e Ambientali.

È attivo in diverse aree di ricerca tra cui, sviluppo e applicazione di metodi di apprendimento automatico per la predizione della funzione genica basati sull'integrazione di dati biomolecolari eterogenei, analisi e sviluppo di metodi supervisionati per l'identificazione automatica di geni e/o trascritti codificanti proteine, analisi di dati di espressione genica per la ricostruzione delle alterazioni dei pathway biomolecolari in malattie tumorali e genetiche, sviluppo di metodi semi-supervisionati per l'analisi di grafi con applicazioni in biologia computazionale, sviluppo e applicazione di metodi per l'analisi di reti di grandi dimensioni in memoria secondaria.

L'attività di ricerca è documentata da 18 articoli pubblicati in riviste internazionali con peer-review e da più di 40 articoli, fra articoli pubblicati su rivista, atti di conferenze internazionali e contributi a volumi collettivi (ad es. Lecture Notes in Computer Science).

Progetti di ricerca:

Ha partecipato alla rete di eccellenza Pattern Analysis, Statistical Modelling and Computational Learning 2 (PASCAL2), come membro dell' unità milanese coordinata da Nicolò Cesa-Bianchi nell'ambito del VII Programma Quadro dell'Unione Europea. Ha partecipato al progetto PUR 2009 Metodi automatici per l'analisi di pattern in ambito biomedico finanziato dall'Università degli Studi di Milano. E' membro dell'unità bioinformatica per l'analisi di dati di DNA microarray nell'ambito di un progetto comune con il Dipartimento di Biologia e Genetica della Facoltà di Medicina e Chirurgia dell'Università degli Studi di Milano e con l'Ospedale Niguarda per la diagnosi e terapia bio-molecolare delle leucemie mieloidi. Collabora a progetti di ricerca comuni con il Dipartimento di Scienze Biomolecolari e Biotecnologie dell'Università degli Studi di Milano. Collabora a progetti di ricerca comuni con l'unità di Targeting Molecolare del Dipartimento di Oncologia Sperimentale della Fondazione IRCCS Istituto Nazionale dei Tumori.

Collaborazioni nazionali e internazionali:

Collaborazioni accademiche:

Collabora con il dipartimento di Informatica della Aristotle University of Thessaloniki (metodi di classificazione multilabel e gerarchici, integrazione e analisi di dati bio-molecolari complessi per la predizione delle funzioni geniche).
È attualmente coinvolto in una collaborazione con l' Institute for Medical Genetics and Human Genetics della Charitè-Universitatsmedizin di Berlino avente come obiettivo la progettazione, sviluppo e applicazione di metodi di machine learning per la predizione di fenotipi patologici umani.
Collabora con il Centre for Systems and Synthetic Biology del dipartimento di Computer Science della Royal Holloway University di Londra per la progettazione e sviluppo di metodi di machine learning che permettano l'integrazione di dati eterogenei rappresentati in forma di grafi e che trovino applicazione in experimenti di biologia computazionale atti a predire la funzione di geni o proteine. Recentemente, sempre con il Centre for Systems and Synthetic Biology, è stata avviata una ulteriore linea di ricerca tesa allo sviluppo di metodi di apprendimento automatico in grado di effettuare predizione di prognosi e suscettibilità a patologie in reti di individui le cui relazioni si basano su dati molecolari complessi quali profili di espressione genica e risultati di analisi biochimico-cliniche. Collabora attivamente con il gruppo di ricerca in bioinformatica e systems biology coordinato dal prof. Von Mering dell'Institute of Molecular Life Sciences dell'università di Zurigo nell'ambito della recente linea di ricerca sullo sviluppo e applicazione di metodi di apprendimento automatico basati su kernel per l'analisi di big data in bioinformatica e biologia computazionale.

Collaborazioni industriali:

Collabora con Italtel per lo sviluppo ed applicazione in area bioinformatica e biomedica di algoritmi di apprendimento automatico ottimizzati per l'esecuzione in piattaforme di calcolo ad alte prestazioni.

Attività editoriale e organizzazione di conferenze/workshop internazionali:

Svolge attività di reviewer per riviste internazionali di bioinformatica (*Bioinformatics*, *BMC Bioinformatics*, *Nucleic Acids Research*, *Artificial Intelligence in Medicine*, *Advances in Bioiformatics*) e per riviste di apprendimento automatico tra cui *Neurocomputing*. Svolge inoltre attività di reviewer per alcune conferenze internazionali tra cui *European Conference on Artificial Intelligence* (ECAI), *Multiple Classifier Systems* (MCS) e *Supervised and Unsupervised Ensemble Methods and their Applications* (SUEMA).
È stato co-chair del workshop internazionale SUEMA 2010 (`http://suema10.di.unimi.it/`) dedicato ai sistemi ensemble supervisionati e non supervisionati ed alle loro applicazioni, svoltosi nell'ambito della conferenza ECML (European Conference on Machine Learning) [19], ed è stato editore di un volume pubblicato dalla Springer contenente i migliori articoli presentati a SUEMA 2010 [21].

E' editore di un libro associato ai lavori presentati al workshop internazionale SUEMA (Supervised and Unsupervised Ensemble Methods and Their Applications) 2010, tenutosi nell'ambito della conferenza internazionale ECML/PKDD 2010 [19]. E' autore di un articolo di review sui sistemi ensemble [21].

E' executive editor dell'International Journal of Neural Networks (ISSN: 2249-2763 print version, E-ISSN: 2249-2771 electronic version, DOI: 10.9735/2249-2763).

Attività didattica:

Attività di docenza in corsi universitari:

A.a. 2015/16:

- Docente del corso di Architetture degli Elaboratori I - Laboratorio, corso di laurea triennale in Informatica, Facoltà di Scienze MFN, Università degli Studi di Milano.

A.a. 2014/15:

- Codocenza nel corso di "Bioinformatica" (I semestre) (corso di laurea magistrale in Informatica, Facoltà di Scienze MFN, Università degli Studi di Milano).
- Docente del corso di Sistemi operativi II (Laboratorio di sistemi operativi), corso di laurea triennale in Informatica, Facoltà di Scienze MFN, Università degli Studi di Milano.

A.a. 2013/14:

- Codocenza nel corso di "Bioinformatica" (I semestre) (corso di laurea magistrale in Informatica, Facoltà di Scienze MFN, Università degli Studi di Milano).
- Docente del corso di Sistemi operativi II (Laboratorio di sistemi operativi), corso di laurea triennale in Informatica, Facoltà di Scienze MFN, Università degli Studi di Milano.

A.a. 2012/13:

- Codocenza nel corso di "Bioinformatica" (I semestre) (corso di laurea magistrale in Informatica, Facoltà di Scienze MFN, Università degli Studi di Milano).
- Docente del corso di "Bioinformatica" (II semestre) per il corso di laurea in Biotecnologie del Farmaco, presso la Facoltà di Farmacia dell'Università degli studi di Milano.

A.a. 2011/12:

- Docente del corso di "Bioinformatica" (II semestre) per il corso di laurea in Biotecnologie del Farmaco, presso la Facoltà di Farmacia dell'Università degli studi di Milano.
- Ha svolto diverse lezioni per il corso di "Biologia Computazionale" (I semestre) per il corso di laurea in Biotecnologie Industriali ed Ambientali, presso la Facoltà di Scienze MFN dell'Università degli studi di Milano.
- Ha tenuto 4 lezioni nel corso di "Bioinformatica" (I semestre) (corso di laurea magistrale in Informatica, Facoltà di Scienze MFN, Università degli Studi di Milano).

A.a. 2010/11:

- Ha tenuto diverse lezioni di "Biologia Computazionale" per il corso di laurea in Biotecnologie Industriali ed Ambientali, presso la Facoltà di Scienze MFN dell'Università degli studi di Milano (`http://homes.di.unimi.it/re/corsobc11.html`).

A.a. 2009/10:

- Ha tenuto diverse lezioni per il corso "Bioinformatica" per la laurea magistrale in Informatica, presso la Facoltà di Scienze MFN dell' Università degli Studi di Milano.

- Ha tenuto diverse lezioni e seminari per il corso "Metodi Bioinformatici" per la laurea magistrale in Biotecnologie Molecolari e Bioinformatica presso la Facoltà di Scienze MFN dell' Università degli Studi di Milano (`http://homes.di.unimi.it/re/corsomb10.html`).

- Ha tenuto diverse lezioni e seminari per il corso "Informatica Avanzata" (Laurea magistrale in Biotecnologie Industriali ed Ambientali , presso la Facoltà di Scienze MFN dell' Università degli studi di Milano).

Ha seguito come relatore o co-relatore varie tesi di laurea magistrale in Informatica, Matematica, Biologia molecolare e Biotecnologie a indirizzo bioinformatico, presso la Facoltà di Scienze MFN dell' Università degli Studi di Milano. Ha seguito come co-relatore una tesi di laurea magistrale in Biotecnologie Mediche e Medicina Molecolare presso la Facoltà di Medicina e Chirurgia dell' Università degli Studi di Milano.

Presentazioni orali a conferenze internazionali:

- Identification of promoter regions in genomic sequences by 1-dimensional constraint clustering, WIRN 2011, Vietri sul mare, Salerno, Italy

- Functional Inference in FunCat through the Combination of Hierarchical Ensembles with Data Fusion Methods. Second International Workshop on Learning from Multi-Label Data, in conjunction with ICML/COLT 2010. June 25, 2010 - Haifa, Israel

- International Symposium on Integrative Bioinformatics, 6th annual meeting: Noise tolerance of multiple classifier systems in data integration-based gene function prediction, 22nd to 24th March 2010, Cambridge, UK.

- Predicting Gene Expression from Heterogeneous Data, Sixth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB), 15-17 October 2009 - Genova (Italy)

- MLD09, ECML workshop on Learning from Multi-Label Data, Sept. 7, 2009: Weighted True Path Rule: a multilabel hierarchical algorithm for gene function prediction. ECML PKDD 2009, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Bled, Slovenia.

- MLSB09, the Third International Workshop on Machine Learning in Systems Biology, Ljubljana, Slovenia,Sept. 5-6 2009. Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction.

- Comparing early and late data fusion methods for gene function prediction. WIRN 2009, May 28-30, 2009, Vietri sul Mare, Salerno, Italy

- Signal Processing in Comparative Genomics. International Workshop on Fuzzy Logic and Applications, WILF 2007. Portofino Vetta - Ruta di Camogli, Genova (Italy) - July 7-10, 2007

Presentazioni orali a conferenze nazionali:

- Analysis of large bio-molecular networks through semi-supervised graph-based learning methods ,PRIN workshop 2015. May 29, 2015, Naples, Italy

- Data fusion based gene function prediction using ensemble methods. Sixth Annual Meeting of the Bioinformatics Italian Society. March 18-20, 2009, Genoa, Italy.

Riconoscimenti:

Terzo Premio di Laurea "Fra Pierluigi Marchesi", Centro S.Giovanni di Dio, Fatebenefratelli I.R.C.C.S., sezione: bioinformatica. Titolo tesi: Identificazione di nuovi geni o isoforme di splicing nel genoma umano mediante analisi di genomica comparata.

Corsi di specializzazione post laurea, nazionali ed internazionali:

- First Lipari International Summer School on Bioinformatics and Computational Biology: "Advances Computational Proteomics: Structure, Imaging and Control", Lipari (Messina) from June 16 to June 23, 2007
- Italian Perl Workshop, Dept. of computer science - Polo Fibonacci - University of Pisa (Italy), 22 to 23 June, 2006

Affiliazione ad associazioni nazionali ed internazionali:

E' membro di BITS (Bioinformatics Italian Society).

Partecipazione a challenge internazionali:

Ha partecipato come membro del gruppo di Bioinformatica e Biologia computazionale (Anacletolab) del dipartimento di Informatica dell'Università degli Studi di Milano (`http://anacletolab.di.unimi.it/`), alla seconda edizione della challenge internazionale Critical Assessment of protein Function Annotation (CAFA) che ha visto più di 100 gruppi di ricerca impegnati nella predizione della funzione di geni e proteine o della loro associazione funzionale con fenotipi patologici umani.
Nei task di predizione più complessi in cui l'obiettivo era la predizione di associazione tra geni o proteine e fenotipi patologici umani, il gruppo Anacletolab ha ottenuto performance competitive (quarto classificato). I risultati prodotti in questa indagine sono stati alla base dell'avvio di una collaborazione stabile con l' Institute for Medical Genetics and Human Genetics della Charitè-Universitatsmedizin di Berlino e sono attualmente sottomessi ,dagli organizzatori di CAFA2, per la pubblicazione su rivista (Genome Biology, `http://www.genomebiology.com/`, impact factor: 10.8).

Indicatori bibliometrici output ricerca in banche dati internazionali:

ELSEVIER SCOPUS (`www.scopus.com`, 05/12/2015):
Author ID: 35305844500
Numero documenti indicizzati : 27 (Article: 15, Conference paper: 11, Editorial: 1)
Periodo di riferimento: 2006 - 2015
N. citazioni: 94
Indice h: 7

Attività di ricerca:

La bioinformatica e l'apprendimento automatico sono le discipline nel cui ambito Matteo Rè ha svolto le proprie attività di ricerca. La formazione di base in Biologia molecolare e bioinformatica, ed il forte interesse e lo studio delle discipline informatiche hanno favorito lo sviluppo di linee di ricerca con un forte grado di interazione fra bioinformatica ed apprendimento automatico.
Schematicamente, l'attività di ricerca si può articolare in due aree principali, a loro volta suddivise in alcune sottoaree:

I. **Bioinformatica**

A) Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di apprendimento automatico supervisionato

A1. Classificazione funzionale di geni e proteine basata su ontologie

A2. Integrazione di sorgenti multiple di dati per la predizione delle funzioni geniche

A3. Integrazione di dati bio-molecolari complessi per la classificazione supervisionata di geni co-espressi

B) Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di apprendimento automatico semisupervisionato per il ranking di nodi in reti di geni e di farmaci

B1. Metodi per la predizione di Cancer Gene Modules basati sull'ordinamento di nodi in reti di geni mediante funzioni di scoring kernelizzate

B2. Sviluppo di metodi per l'integrazione di reti di farmaci basati su proiezioni di reti bipartite per il riposizionamento di farmaci in nuove classi terapeutiche

B3. Analisi di big data in biologia computazionale e bioinformatica. Predizione multi specie della funzione di geni e proteine mediante analisi ed integrazione di reti biomolecolari di grandi dimensioni.

C) Altre attività di ricerca in ambito bioinformatico

C1. Metodi per l'identificazione di regioni genomiche e trascritti codificanti proteine

C2. Analisi di dati di DNA microarray per lo studio della leucemia mieloide

II. **Apprendimento automatico**

A) Sviluppo e analisi di metodi di ensemble multiclasse, multietichetta e multi-path per problemi di classificazione gerarchica

B) Metodi semi-supervisionati per l'ordinamento e la classificazione di nodi in grafi

C) Metodi di ensemble per l'integrazione di dati eterogenei

D) Metodi semi-supervisionati per l'analisi di grafi di grandi dimensioni mediante l'utilizzo di kernel definiti su grafo

E) Sviluppo di librerie software per la realizzazione di esperimenti di Machine Learning

F) Pubblicazione di dataset

Nel seguito sono descritte più in dettaglio le attività di ricerca relative alle due aree principali.

I. Bioinformatica :

Le attività di ricerca in bioinformatica di Matteo Rè sono caratterizzate dallo sviluppo ed applicazione di metodi ed algoritmi di apprendimento automatico per l'estrazione di conoscenza biologica dall'analisi di dati bio-molecolari generati da bio-tecnologie high-throughput[13]. Nel seguito sono riassunte schematicamente le principale aree e linee di ricerca.

A) **Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di apprendimento automatico supervisionato.**
L'attività di ricerca si è concentrata soprattutto sullo sviluppo di metodi bioinformatici basati su ensemble di learning machine. In particolare si possono distinguere tre linee di ricerca principali:

A1. Classificazione funzionale di geni e proteine basata su ontologie

A2. Integrazione di sorgenti multiple di dati per la predizione delle funzioni geniche

A3. Integrazione di dati bio-molecolari complessi per la classificazione supervisionata di geni co-espressi

I.A1: Classificazione funzionale di geni e proteine basata su ontologie
Nel contesto delle problematiche legate alla predizione delle funzioni di geni/proteine con metodi computazionali, l'analisi dei grafi della Gene Ontology (GO) e degli alberi di FunCat del MIPS di Monaco, tramite cui sono strutturate le relazioni fra le classi funzionali di geni, sono di grande rilevanza in ambito bioinformatico. Per la classificazione strutturata delle classi funzionali dei geni è essenziale introdurre procedure automatiche per l'associazione dei geni alle classi funzionali e a diverse tipologie di dati bio-molecolari. A tal fine si sono sviluppati metodi ed algoritmi tramite cui è possibile selezionare classi funzionali di geni/proteine correlati a specifici problemi biologici, effettuare il pre-processing di dati bio-molecolari complessi e multi-view, e supportare lo sviluppo di metodi di classificazione gerarchica di geni basati sulle tassonomie della Gene Ontology (GO) e di FunCat. La predizione della funzione dei geni è un problema di classificazione multiclasse e multietichetta complesso caratterizzato da una strutturazione gerarchica delle classi. Nel corso degli ultimi 2 anni si sono sviluppati metodi di classificazione gerarchici per la predizione delle classi funzionali di geni/proteine, basati su ensemble di learning machine strutturate ad albero. In particolare per FunCat si sono sviluppati metodi di ensemble basati sulla "true path rule" (TPR) che governa sia FunCat che la GO[30] e metodi bayesiani cost-sensitive per la riconciliazione probabilistica dell'output dei base learner.
Entrambi i metodi, benchè derivino da impostazioni teoriche ed euristiche differenti, hanno mostrato risultati comparabili, almeno quando un'unica sorgente di dati biomolecolari viene utilizzata per la classificazione dei geni a livello dell' intero genoma e dell' intera tassonomia FunCat[28]. L'estensione del metodo basato sulla "true path rule" alle tassonomie basate su DAG (i.e. la GO e la Human Phenotype Ontology) è stata recentemente completata. L'applicazione dei nuovi metodi sviluppati in problemi rilevanti in biologia computazionale ha prodotti risultati incoraggianti pubblicati in[22, 23].
Il problema della predizione delle funzioni geniche mediante l'applicazione di metodi di apprendimento supervisionato ha diverse caratteristiche distintive quali l'esistenza di relazioni gerarchiche tra le classi funzionali dei geni, la necessità di integrare sorgenti multiple di dati biomolecolari e un elevato sbilanciamento tra esempi positivi e negativi. In letteratura sono descritti molti metodi sviluppati appositamente per affrontare questi problemi separatamente ma una valutazione sistematica della sinergia tra approcci di classificazione supervisionata gerarchici e multietichetta, di metodi di integrazione di dati eterogenei e di approcci cost-sensitive che permettano di affrontare efficacemente lo sbilanciamento delle classi da predire, non è al momento disponibile. Sono stati quindi effettuati dei test che hanno permesso di dimostrare tale effetto sinergico e che suggeriscono come, nell'ambito di esperimenti di predizione della funzione genica, un ruolo centrale sia svolto anche da

opportune tecniche si selezione degli esempi negativi[9].
In prospettiva, integrando le linee di ricerca sui metodi di ensemble gerarchici con quella per l'integrazione di sorgenti multiple di dati[9, 29], si prevedono applicazioni alla predizione delle funzioni delle proteine di S.cerevisiae, C. elegans, A. thaliana e M. musculus.
Seguendo un approccio basato su reti di Hopfield è stato sviluppato e validato sperimentalmente un classificatore in grado di apprendere efficacemente anche in presenza dell'elevato tasso di sbilanciamento tra positivi e negativi che costituisce un elemento caratterizzante in problemi di predizione della funzione genica. Le performance predittive di questo classificatore sono state testate estensivamente e confrontate con metodi allo stato dell'arte ottenendo risultati incoraggianti[4].

I.A2: Integrazione di sorgenti multiple di dati per la predizione delle funzioni geniche
Singole sorgenti di dati biomolecolari sono in genere predittive solo per alcune classi funzionali, mentre possono risultare totalmente non informative per altre, poichè ogni sorgente di dati cattura solo alcune delle caratteristiche funzionali dei geni e dei prodotti genici. Per questa ragione l'integrazione di diverse sorgenti di dati è un problema centrale in bioinformatica.
A questo fine si sono sviluppati approcci basati su ensemble di learning machine [12, 35, 34, 32], mostrando che anche metodi relativamente semplici come la votazione maggioranza o i decision template possono ottenere risultati comparabili con lo stato dell'arte[10, 31]. Si è inoltre mostrato che i metodi di ensemble sono in grado di tollerare anche relativamente elevati livelli di rumore nei dati, senza una significativo deterioramento delle prestazioni[11].
Recentemente l'analisi di dati biologici rappresentati sotto forma di grafo sta acquisendo sempre maggior popolarità. Tale trend è facilmente motivabile considerando che le funzioni biologiche sono estremamente complesse e, di conseguenza, solo di rado avviene che una determinata funzione sia sotto il diretto controllo di un unico gene o di un'unica proteina. Le reti biologiche costituiscono quindi il mezzo di elezione per l'analisi delle complesse interazioni tra geni che, in caso di alterazione, possono anche essere all'origine di importanti patologie. In questo contesto l'analisi di reti di interazioni geniche permettono di predire l'associazione funzionale tra geni e patologie. Nonostante la provata efficacia di metodi che predicono associazioni tra geni e malattie basati sull'analisi di reti geniche, in letteratura l'integrazione di reti geniche derivate da dati eterogenei ha ricevuto poca attenzione. E' stata quindi effettuata un'estensiva analisi dell'impatto dell'integrazione di reti eterogenee in task di predizione di associazioni tra geni e malattie. Oltre ai tipi di dati classicamente utilizzati in questo tipo di analisi per la costruzione delle reti, sono stati considerate anche reti derivanti da misure di similarità semantica tra i geni basate su dati estratti da letteratura biomedica. I risultati ottenuti, competitivi con lo stato dell'arte, sono stati pubblicati in [1].

I.A3: Integrazione di dati bio-molecolari complessi per la classificazione supervisionata di geni co-espressi
La predicibilità di classi di geni co-espressi dalle regioni regolatorie non codificanti del DNA è un problema aperto che può fornire indirettamente informazioni sui siti di legame dei fattori di trascrizione e sui motivi regolatori delle regioni non codificanti del DNA. Dal punto di vista dell'apprendimento automatico è un problema di classificazione particolarmente rilevante per la complessità delle regioni regolatorie e per l'ambiguità dei motivi regolatori stessi, e per la necessità di integrare dati di sequenza (le regioni regolatorie non codificanti del DNA) e dati di espressione genica (per l'individuazione di classi di geni co-espressi). Gli approcci attualmente disponibili, seppur efficaci, come dimostrato da diversi lavori recentemente pubblicati in letteratura, non tengono conto della sempre crescente quantità di dati prodotta da tecnologie high-throughput

applicate nell'ambito di progetti genomici. Questa limitazione può essere risolta mediante l'applicazione di tecniche di predizione basate sull'integrazione di dati eterogenei.

Grazie a questi metodi è possibile integrare, oltre ai dati riguardanti le sequenze regolatorie site nei promotori e i pattern di espressione, informazioni inerenti l'occorrenza di pattern epigenetici complessi (come le modificazioni istoniche) o pattern di costrizioni, agenti nel corso dell'evoluzione, tesi alla conservazione di alcuni segnali regolatori presenti nei promotori, assicurando un'incremento nella specificificità delle predizioni prodotte senza ridurre la sensibilità complessiva.

Tra le varie tecniche di integrazione disponibili i metodi ensemble sono particolarmente adatti per l'applicazione in questa particolare area di ricerca. Ciò è dovuto principalmente al fatto che, nei sistemi ensemble, l'integrazione dei dati avviene a livello delle decisioni prodotte dai classificatori componenti, e questo permette l'integrazione di dati strutturalmente diversificati (ad esempio grafi che esprimono reti di interazioni tra proteine regolatorie o vettori numerici che esprimono, su base posizionale, la forza delle pressioni evolutive agenti sulle regioni promotore dei geni). L'applicazione di metodi ensemble per la predizione di gruppi di geni coespressi basata sull'integrazione di informazioni inerenti ai pattern di modificazione istonica ed alla conservazione evolutiva di motivi regolatori oltre ai classici dati inerenti ai motivi regolatori ed ai pattern di espressione genica ha prodotto risultati incoraggianti[33, 13].

B) **Analisi sviluppo ed applicazione in ambito bioinformatico di metodi di apprendimento automatico semisupervisionato per il ranking di nodi in reti di geni e di farmaci.**

L'attività di ricerca si è concentrata soprattutto sullo sviluppo di metodi bioinformatici per l'ordinamento di nodi in grafi e per l'integrazione di dati eterogenei strutturati in forma di grafi. In particolare si possono distinguere due linee di ricerca principali:

B1. Metodi per la predizione di Cancer Gene Modules e per la predizione delle funzioni geniche basati sull'ordinamento di nodi in reti di geni mediante funzioni di scoring kernelizzate

B2. Sviluppo di metodi per l'integrazione di reti di farmaci basati su proiezioni di reti bipartite per il riposizionamento di farmaci in nuove classi terapeutiche

I.B1: Metodi per la predizione di Cancer Gene Modules basati sull'ordinamento di nodi in reti di geni e sull'utilizzo di funzioni di scoring kernelizzate

Le variazioni nei pattern di espressione genica sono state ampiamente utilizzate per la caratterizzazione dal punto di vista molecolare e prognostico di varie classi di tumori. Un'analisi comparativa di tali pattern in vari tipi e sottotipi di tumore ha portato all'identificazione di set di geni (Cancer Gene Modules o CM) che agiscono in sincronia per la realizzazione di specifici processi biologici. L'utilizzo di questi set di geni ha permesso di caratterizzare le varie classi e sottoclassi di tumori rispetto alle combinazioni di moduli attivati o disattivati. Sfortunatamente i CM sono stati definiti unicamente sulla base di fluttuazioni nei pattern di espressione genica ignorando altri tipi di dati biomolecolari (ad esempio interazioni tra proteine). E' quindi di estremo interesse l'utilizzo di dati biomolecolari eterogenei per la predizione dell'appartenenza di geni ai Cancer Gene Modules.

La predizione dell'appartenenza dei geni a specifici CM è stata formalizzata come un problema semisupervisionato di ranking di nodi in un grafo indiretto. Il ranking è ottenuto tramite opportune funzioni di score basate su kernel, estendendo un approccio recentemente proposto da Borgwardt e collaboratori[1]: qualsiasi kernel può essere utilizzato, ma in questo contesto i graph kernel ed in particolare i random walk kernel

[1] *Lippert C., Ghahramani Z., Borgwardt K.M. Gene function prediction from synthetic lethality networks via ranking on demand. Bioinformatics. 2010 Apr 1;26(7):912-8.*

sono in grado di catturare opportunamente le relazioni funzionali fra i geni codificate nella topologia della rete genica. In particolare la matrice kernel viene construita a partire dalla matrice di adiacenza del grafo e successivamente utilizzata da score function in grado di generare un ranking a partire da un nucleo di geni la cui appartenenza al CM è nota a priori.
I risultati ottenuti su un ampio set di CM [25, 5] hanno dimostrato le potenzialità di questo approccio producendo risultati comparabili (in alcuni casi migliori) di quelli ottenuti mediante l'applicazione di metodi allo stato dell'arte per la risoluzione del problema considerato e tempi di calcolo inferiori di uno o due ordini di grandezza (a seconda delle funzioni di scoring considerate). I metodi sviluppati sono stati applicati anche a problemi di predizione della funzione genica con risultati competitivi con lo stato dell'arte [6].

I.B2: Sviluppo di metodi per l'integrazione di reti di farmaci basati su proiezioni di reti bipartite per il riposizionamento di farmaci in nuove classi terapeutiche

Lo sviluppo di nuovi farmaci è un processo costoso e fortemente soggetto a possibili fallimenti. Negli ultimi anni un nuovo paradigma di ricerca farmacologica noto come "Drug Repositioning" (riposizionamento di farmaci in classi terapeutiche differenti da quelle per cui erano stati inizialmente sviluppati) sta emergendo in quanto è in grado di ridurre i costi di sviluppo ed i tempi necessari all'immissione di nuovi farmaci sul mercato che richiedono, tipicamente, 10-15 anni ed investimenti che superano il miliardo di dollari. In linea di principio il problema del riposizionamento di farmaci può essere affrontato mediante le tecniche sviluppate nella linea di ricerca I.B1 (mediante ranking di farmaci rappresentati sotto forma di nodi di una rete di similarità ed utilizzo di funzioni di scoring kernelizzate). A differenza di altri problemi bioinformatici che coinvolgono l'utilizzo di un'unica rete, il problema del riposizionamento di farmaci coinvolge l'utilizzo simultaneo di numerose reti bipartite in cui troviamo i farmaci associati ad un secondo gruppo di nodi (ad es. proteine con cui il farmaco interagisce o malattie o geni i cui livelli di espressione risultano alterati in seguito alla somministrazione del farmaco). E' stato sviluppato un metodo che permette l'integrazione di queste reti eterogenee basato sulla proiezione di ogni rete bipartita in una rete omogenea composta unicamente da farmaci in cui un arco tra due nodi esprime la condivisione di relazioni con i nodi appartenenti al secondo set di nodi nella rete bipartita proiettata. Le reti farmacologiche ottenute vengono quindi integrate in un'unica rete in cui è possibile applicare le funzioni di scoring kernelizzate utilizzando, come set di positivi, i farmaci appartenenti a set definiti sulla base di categorie terapeutiche note. Risultati sperimentali condotti su 1300 farmaci approvati dalla Food and Drugs Administration (FDA) americana dimostrano l'efficacia dei metodi sviluppati [24, 7, 3].

I.B3: Analisi di big data in biologia computazionale e bioinformatica. Predizione multi specie della funzione di geni e proteine mediante analisi ed integrazione di reti biomolecolari di grandi dimensioni

L'applicazione di metodiche di apprendimento automatico e lo sviluppo di nuove tecniche in grado di modellare efficacemente fenomeni biologici al fine di estrarre nuova conoscenza da dati molecolari e clinici complessi ha acquisito sempre maggior popolarità nell'ultimo decennio.
Tra le tecniche che hanno avuto maggior successo in problemi quali la predizione delle funzioni geniche o il coinvolgimento di specifici geni nell'insorgenza e progressione di malattie ad elevato impatto sociale possono sicuramente essere annoverati i metodi kernel ossia metodi in grado di modellare relazioni tra oggetti in spazi alto dimensionali permettendo di esprimere, mediante opportune tecniche di rappresentazione, misure di similarità tra coppie di oggetti che possono essere utilizzate per l'addestramento di classificatori.
Nonostante la loro provata efficacia e precisione l'utilizzo di metodi di apprendimento basati su kernel in biologia computazionale e bioinformatica si scontra con alcune importanti limitazioni dei metodi kernel.

In primo luogo la maggioranza dei metodi kernel applicati in bioinformatica segue un paradigma di apprendimento supervisionato in cui il classificatore richiede una fase di addestramento dipendente dalla disponibilità di quantità rilevanti di esempi annotati con etichette di alta qualità. Nelle scienze di area biomedica tale disponibilità non è scontata (basti pensare alle difficoltà insite nel processo di diagnosi o agli elevati costi associati ad esperimenti di laboratorio tesi a scoprire il meccanismo di azione di singoli geni). Un secondo problema che limita l'applicazione di metodi kernel in bioinformatica è costituito dalla loro scarsa scalabilità, dovuta principalmente alle complessità temporali e spaziali associata al calcolo della matrice di Gram che esprime le relazioni di similarità tra tutti gli esempi considerati nel problema di classificazione.
Grazie all'utilizzo di moderne biotecnologie un singolo esperimento pu produrre informazioni su milioni di molecole e nell'ultima release della banca dati di riferimento delle proteine è possibile trovare informazioni su milioni di proteine provenienti da migliaia di specie. In queste condizioni, l'applicazione di metodi kernel non è possibile. E' stata recentemente avviata una linea di ricerca che ha permesso di estendere il framework algoritmico delle funzioni di score kernelizzate in modo da renderlo applicabile a dataset di grandi dimensioni, come quelli generalmente analizzati in aree di ricerca quali l'analisi delle reti sociali. Grazie alle tecniche sviluppate è possibile effettuare predizioni in reti biomolecolari composte da milioni di esempi utilizzando kernel definiti su grafo (ad esempio random walk kernel). I problemi di scalabilità dei metodi kernel hanno limitato il loro utilizzo in problemi di predizione della funzione di geni e proteine utilizzando più di un organismo (esistono organismi in cui il numero di proteine è dell'ordine delle decine di migliaia e il numero degli organismi per i quali è disponibile un catalogo dei geni non completamente annotato dal punto di vista funzionale è dell'ordine delle migliaia). Le tecniche sviluppate in questa linea di ricerca hanno permesso di realizzare, per la prima volta, un esperimento di predizione della funzione proteica di più di 300 organismi [2].

C) **Altre attività di ricerca in ambito bioinformatico**
Si possono distinguere due ulteriori linee di ricerca:

C1. Metodi per l'identificazione di regioni genomiche e trascritti codificanti proteine

C2. Analisi di dati di DNA microarray per lo studio della leucemia mieloide

I.C1: Metodi per la classificazione di regioni genomiche e trascritti codificanti proteine
L'identificazione dei geni codificanti proteine nel genoma umano à un problema complesso basato sull'analisi di proprietà composizionali del DNA e sui pattern di conservazione osservabili mediante confronto dei genomi di diversi organismi. Nonostante la sequenza completa del genoma umano sia disponibile da quasi undici anni ad oggi non à ancora disponibile il catalogo completo dei geni umani. Ciò è dovuto all'esistenza di molti geni aventi caratteristiche peculiari (geni senza introni, geni particolarmente corti, geni aventi composizione nucleotidica non standard). Un possibile approccio per l'identificazione di questi geni "nascosti" è costituito dalle metodiche di genomica comparata che coinvolgono informazioni derivanti dal confronto delle sequenze genomiche di più organismi. Sono stati sviluppati metodi di localizzazione dei geni non annotati basati sulla valutazione dei pattern di pressione evolutiva caratterizzanti le sequenze codificanti proteine. Tali metodi[18] sono stati resi disponibili per l'utilizzo da parte della comunità scientifica mediante l'applicazione di tecnologie GRID[17]. Nonostante la dimostrata capacità dei metodi di identificazione delle regioni genomiche codificanti proteine basati sull'analisi dei pattern evolutivi, essi sono affetti da gravi limitazioni nell'analisi di sequenze relativamente corte (60-120 nucleotidi, corrispondenti a 20-40 aminoacidi). Di fatto i metodi di valutazione del potenziale codificante ad oggi disponibili si limitano ad effettuare l'analisi solo per sequenze di lunghezza maggiore di 100 nucleotidi. Sono stati quindi sviluppati metodi che permettono l'analisi di regioni genomiche più corte mediante l'applicazione di metodi di digital signal processing che valutano la forza delle pressioni evolutive agenti sulle regioni conservate

in termini di periodicità di segnale associata ai pattern di sostituzione nucleotidica[15, 36]. I metodi proposti basati su analisi dei segnali sono stati integrati con gli approcci presenti in precedenza in letteratura mediante l'utilizzo di metodi di classificazione supervisionata addestrando modelli calibrati per diverse classi di lunghezza delle sequenze i quali, sfruttando contemporaneamente pattern di composizione locale e di sostituzine permettono di ottenere classificazioni allo stato dell'arte[14]. La validazione del metodo proposto à stata effettuata considerando parte del genoma umano ed utilizzando solo due specie (H. *sapiens* e M. *musculus*) per l'analisi dei pattern evolutivi. E' attualmente in corso l'estensione dell'esperimento all'intero genoma umano. Una possibile estensione della linea di ricerca, ad oggi assente in letteratura, sarebbe costituita dall'utilizzo , durante la stima del potenziale codificante delle regioni genomiche conservate, tra più di due organismi. Questa linea di ricerca è stata recentemente potenziata mediante lo sviluppo ed applicazione di metodi per la caratterizzazione funzionale del prodotto dei geni (proteine) e per la predizione di regioni promotore (regioni che regolano l'attivazione ed il silenziamento dei geni in risposta a particolari condizioni intracellulari). I risultati preliminari ottenuti sono incoraggianti [26, 27].

I.C2: Analisi di dati di DNA microarray per lo studio della leucemia mieloide

Tale attività si colloca nell'ambito di un accordo-quadro tra l'Universit degli Studi di Milano (Dipartimenti di Genetica di Medicina e Dipartimento di Scienze dell'Informazione,DSI di Scienze MFN) e l'Ospedale di Niguarda, per l'elaborazione di dati di espressione genica relativi a pazienti affetti da patologie tumorali ed in particolare per lo studio della leucemia mieloide. Le attività di studio e ricerca permetteranno di applicare a dati reali, ottenuti tramite la piattaforma bio-tecnologica Affymetrix di Niguarda, metodi computazionali per il pre-processing e il controllo di qualità dei dati di espressione genica, per l'analisi e rilevazione di geni differenzialmente espressi, per la ricerca di geni correlati a patologie tumorali e per l'analisi di reti funzionali di geni, con applicazioni rilevanti in ambito bio-medico. In tale contesto si sono sviluppati metodi computazionali per la valutazione dell'affidabilità di cluster di geni individuati tramite algoritmi gerarchici[16].
L'applicazione di metodi di analisi statistica nello studio dei pattern di espressione genica alterati osservabili in cellule che innescano la progressione della leucemia mieloide ha permesso di identificare un gene chiave nel processo di progressione della malattia (Wnt) e di chiarire il meccanismo molecolare (alterato) con cui esso agisce nella patologia[8].

II. Apprendimento automatico :

L' attività di ricerca in apprendimento automatico, benchè collegata con la bioinformatica, presenta linee di lavoro disciplinari autonome, schematizzabili in analisi e progettazione di metodi di ensemble per problemi multi-classe gerarchici, sviluppo di metodi semisupervisionati per l'ordinamento e classificazione di nodi in grafi, analisi e svilppo di metodi ensemble per l'integrazione di dati eterogenei e sviluppo di librerie software di apprendimento automatico. Nel seguito sono riassunte schematicamente le principale aree e linee di ricerca.

A) Sviluppo e analisi di metodi di ensemble multiclasse, multietichetta e multi-path per problemi di classificazione gerarchica.

B) Metodi semisupervisionati per l'ordinamento e la classificazione di nodi in grafi.

C) Metodi di ensemble per l'integrazione di dati eterogenei.

D) Metodi semi-supervisionati per l'analisi di grafi di grandi dimensioni mediante l'utilizzo di kernel definiti su grafo

E) Sviluppo di librerie software per il supporto ad esperimenti di Machine Learning.

II.A: Sviluppo e analisi di metodi di ensemble multiclasse, multietichetta e multi-path per problemi di classificazione gerarchica

Il problema della classificazione funzionale dei geni ha stimolato la ricerca e sviluppo di algoritmi di classificazione multiclasse (le classi funzionali dei geni sono dell'ordine delle centinaia o migliaia, a secondo dell'ontologia di rifrimento considerata), multietichetta (un gene puo' appartenere a più classi) e multi-path (le classi sono strutturate secondo alberi o DAG). Gli algoritmi, benchè sviluppati per rispondere ad un problema bioinformatico, hanno una valenza più ampia, e pongono problematiche interessanti anche dal punto di vista dell'apprendimento automatico[30, 28].

In questa linea di ricerca sono state sviluppate estensioni di un algoritmo di ensemble gerarchico precedentemente proposto in letteratura. Il True Path Rule Ensemble (TPR)[2] è un algoritmo basato sulla logica di annotazione che governa le principali ontologie pubbliche per la classificazione funzionale dei geni, GO e FunCat. In queste ontologie le classi funzionali sono composte da centinaia o migliaia di nodi organizzati in forma di albero o di DAG. La regola fondamentale che garantisce la consistenza delle annotazioni dei geni è nota come "True Path Rule" e può essere riassunta come segue: "L'annotazione di un gene per una classe nella gerarchia è automaticamente trasferita a tutte le classi ancestor mentre un gene non annotato per una determinata classe non può essere annotato in nessuno dei discendenti della classe". Il metodo ensemble TPR è ispirato alla True Path Rule e realizza due flussi di informazione asimmetrici che percorrono la struttura dell'ensemble gerarchico: le predizioni positive dei classificatori componenti influenzano in maniera ricorsiva le predizioni dei classificatori associati ai nodi ancestor mentre le predizioni negative influenzano le predizioni dei classificatori associati ai nodi discendenti della classe considerata.

Questo approccio ha il vantaggio di produrre predizioni che sono immediatamente utilizzabili per produrre nuove annotazioni in quanto compatibili con la TPR che assicura la consistenza delle annotazioni funzionali. I TPR ensemble sono in grado di ottenere risultati allo stato dell'arte ma non tengono conto dell'elevato sbilanciamento tra esempi positivi e negativi che si osserva comunemente nelle classi funzionali più distanti dalla radice delle ontologie. In [30] è stata proposta una variante cost-sensitive del metodo ensemble TPR (weighted TPR o TPR-w) che è in grado di controllare efficacemente mediante un unico parametro il bilanciamento tra precisione e recall. Tale variante permette di ottenere incrementi significativi di performance a livello globale (ossia considerando l'intera ontologia) e, al tempo stesso, di preservare elevati valori di precisione a livello delle foglie, prerequisito fondamentale per permettere l'utilizzo del metodo in esperimenti su vasta scala che richiedano la verifica sperimentale delle predizioni, dati gli alti costi ad essa associati. Il metodo TPR-w è stato inoltre utilizzato nei test descritti in [9] nei quali è stato dimostrato un effetto sinergico in grado di incrementare ulteriormente le prestazioni quando ad ogni nodo dell'ontologia al posto di classificatori addestrati mediante fonti dati singole vengono utilizzati ensemble che integrino fonti dati eterogenee.

Data l'importanza delle ontologie funzionali strtutturate in forma di grafi diretti aciclici (DAG), tra cui si possono annoverare la Gene Ontology(GO), che registra associazioni tra geni e funzioni geniche, e la Human Phenotype Ontology (HPO) che registra associazioni tra geni e fenotipi patologici, l'algoritmo TPR è stato esteso in modo da permetterne l'utilizzo nella predizione di a classi strutturate in forma di DAG. SOno attualmente in corso esperimenti estesi per valutare le performance delle varianti TPR-DAG sviluppate in diversi problemi rilevanti in bioinformatica. I risultati preliminari ottenuti sono incoraggianti e sono stati presentati in conferenze internazionali[23]. I risultati ottenuti evidenziano come gli algoritmi sviluppati siano efficaci sia in problemi di predizione della funzione genica[23] che in problemi di predizione dell'associazione tra geni e fenotipi patologici[22].

II.B: Metodi semi-supervisionati per l'ordinamento e la classificazione di nodi in grafi

La rappresentazione di dati complessi in forma di grafi è divenuta prassi comune in molte aree di ricerca. Tale tipo di rappresentazione, inoltre, è di estrema utilità in molte aree di applicazione tra cui biologia computazionale,

[2] *G. Valentini, True Path Rule hierarchical ensembles for genome-wide gene function prediction, IEEE ACM Transactions on Computational Biology and Bioinformatics, vol.8 n.3 pp. 832-847, 2011.*

analisi di reti sociali e sviluppo di applicazioni per il web.
In apprendimento automatico esistono numerose aree di ricerca dedicate allo sviluppo di algoritmi in grado di predire le etichette dei nodi di grafi. In questa linea di ricerca è stato considerato un problema differente in cui l'obiettivo non è la predizione delle etichette dei nodi ma, piuttosto, l'ordinamento reciproco dei nodi rispetto ad una caratteristica di interesse (i.e. potenziale appartenenza di nodi ad una via metabolica o ad un dato processo patologico in reti di geni). Data l'elevata dimensionalità delle reti biologiche sono stati sviluppati metodi di ordinamento efficienti basati su funzioni di score kernelizzate. Tali funzioni, inoltre, permettono di ottenere un ranking di nodi non etichettati a partire da un nucleo (anche molto ristretto) di nodi positivi risolvendo un problema centrale in biologia computazionale: l'estrema difficoltà nella definizione di esempi negativi. I metodi sviluppati sono stati confrontati con altri metodi allo stato dell'arte ed hanno ottenuto risultati comparabili in termini di precisione a livelli fissati di recall. Valutazioni empiriche della complessità computazionale hanno dimostrato che i metodi proposti sono in grado di ridurre i tempi di calcolo di 1-2 ordini di grandezza rispetto ad altri metodi di node-rnking allo stato dell'arte [25, 3, 5, 6, 1].

II.C: Metodi di ensemble per l'integrazione di dati eterogenei

Il problema della predizione della classe funzionale dei geni è caratterizzato, dal punto di vista dell'apprendimento automatico, dall'esistenza di relazioni gerarchiche tra le classi funzionali, dalla presenza di più sorgenti di dati biomolecolari e da un'elevato sbilanciamento tra positivi e negativi. In questa linea di ricerca sono state valutate le potenzialità di sistemi di integrazione di dati eterogenei basati su sistemi ensemble. I sistemi ensemble sono particolarmente attraenti per questo tipo di problema in quanto essi operano l'integrazione a livello delle decisioni dei classificatori componenti permettendo di integrare dati strutturalmente differenti (ad es. dati in forma vettoriale, dati codificati in forma di grafo). I risultati ottenuti da questa linea di ricerca hanno permesso di dimostrare che metodi di integrazione relativamente semplici quali i sistemi ensemble basati su media pesata e i templati di decisione sono competitivi con metodi allo stato dell'arte[10]. E' stato inoltre possibile dimostrare per questi metodi un buon livello di tolleranza a dati rumorosi[11], osservazione che era stata precedentemente riportata in letteratura unicamente per metodi basati sulla fusione (pesata e non pesata) di kernel. E' stata infine proposta una metodolgia integrata per la classificazione gerarchica multi-label e multi-view basata sull'analisi delle sinergie tra metodi di ensemble gerarchici per la classificazione multietichetta, metodi di integrazione di dati eterogenei e metodi cost-sensitive in grado di gestire dati fortemente sbilanciati fra le classi[9, 33, 1].

II.D: Metodi semi-supervisionati per l'analisi di grafi di grandi dimensioni mediante l'utilizzo di kernel definiti su grafo

Le tecniche di apprendimento automatico sviluppate in questa linea di ricerca fanno uso di recenti tecnologie che permettono di processare grafi di grandi dimensioni. Invece di utilizzare un approccio classico in cui il grafo, usualmente descritto per mezzo di una matrice di adiacenza pesata, deve essere caricato interamente in RAM , il grafo viene serializzato si disco in file aventi una struttura tale da permetterne il processamento vertice per vertice e minimizzare, al contempo, il numero degli accessi casuali in lettura e scrittura necessari per il l'analisi dell'intero grafo. Queste tecniche di elaborazione di grafi di grandi dimensioni sono state proposte recentemente e non sono mai state utilizzate in biologia computazionale.
Elemento ancor più interessante, dal punto di vista della ricerca in apprendimento automatico, è il fatto che metodi di questo tipo non sono mai stati utilizzati per la realizzazione di algoritmi di apprendimento basati su kernel. Infatti il trend attuale nell'area di ricerca (molto attiva) di miglioramento delle caratteristiche di scalabilità dei metodi kernel, è la ricerca di soluzioni in grado di approssimarne la matrice di Gram, il cui calcolo rappresenta indubbiamente uno step computazionalmente intensivo.
Grafi di grandi dimensioni una volta serializzati su disco vengono processati secondo una logica vertice centrica senza garanzie riguardanti l'ordine di processamento dei vertici. Questo rende difficile (in alcuni casi impossibile) implementare algoritmi di apprendimento automatico pre esistenti secondo questa logica di programmazione. Al

meglio delle nostre conoscenze nessun metodo di apprendimento automatico basato su kernel definito si grafo è mai stato espresso in logica di calcolo vertice centrica. Il nostro contributo si è focalizzato sullo sviluppo di algoritmi che permettono l'utilizzo del kernel in modo locale (rispetto ai vertici del grafo) e senza richiedere la presenza in memoria dell'intera matrice di Gram ma solo del vicinato diretto del vertice considerato in un dato momento dall'algoritmo di classificazione. Questo rende possibile non solo il processamento di grafi contenenti milioni di vertici e di applicare i metodi sviluppati in biologia computazionale ma permette, inoltre di utilizzare metodi kernel esatti (non soluzioni approssimate). Metodi sviluppati in questa linea di ricerca sono stati recentemente pubblicati in[2] e si stanno sviluppando varianti che permettano di utilizzare in ottica locale non solo il random walk kernel ma altri tipi di kernel applicabili problemi di apprendimento in cui le relazioni tra esempi siano modellate in forma di grafo.

II.E: Sviluppo di librerie software di machine learning

L'attività di ricerca nell'ambito dei metodi di apprendimento automatico è stata sempre accompagnata da attività di progettazione ed implementazione di librerie software: in particolare i nuovi metodi di ensemble realizzati durante le attività di ricerca, insieme con altri metodi classici pubblicati in letteratura sono stati implementati ed estesi in modo da agevolare la realizzazione di esperimenti di classificazione basati su integrazione di dati eterogenei su vasta scala (whole genome/whole ontology level). Sono in corso di sviluppo librerie software per la classificazione gerarchica e per l'integrazione di dati eterogeni basate su metodi di ensemble.
Le librerie sono state realizzate in linguaggio R (con alcune routine scritte in linguaggio C per ragioni di efficienza) e rilasciate sotto forma di package scaricabili dal repository pubblico CRAN (`http://cran.r-project.org`). Nella realizzazione di queste librerie è stata prestata particolare attenzione all'utilizzo di implementazioni efficienti in modo da permettere la loro applicazione in esperimenti su vasta scala.
Sono stati recentemente pubblicati su CRAN i seguenti package R:

- **PerfMeas.** Libreria per il calcolo di misure di performance in esperimenti di apprendimento automatico. Implementa le principali metriche per la valutazione delle performance di metodi di apprendimento automatico, tra cu AUC (area under the curve), precisione, sensibilità , F-score, area sotto la curva precision-recall, etc. Può essere utilizzata per il calcolo delle performance sia in esperimenti di ordinamento che di classificazione. Supporta esperimenti su vasta scala con numero elevato di classi. Supporta inoltre misure gerarchiche specifiche per problemi di classificazione gerarchici recentemente proposte in letteratura. Disponibile nel repository pubblico di pacchetti R CRAN all'indirizzo:
  `http://cran.r-project.org/web/packages/PerfMeas/index.html`.

- **NetPreProc.** Libreria per il preprocessing e normalizzazione di dati strutturati in forma di grafo. Nella versione corrente implementa numerosi schemi di normalizzazione (alcuni proposti in esperimenti su vasta scala presenti in letteratura), semplificando la realizzazione di esperimenti che coinvolgano il confronto con metodi precedentemente pubblicati. Disponibile nel repository pubblico di pacchetti R CRAN all'indirizzo:
  `http://cran.r-project.org/web/packages/NetPreProc/index.html`.

- **BioNetData.** Collezione di dataset di tipo biomolecolare e chimico strutturati in forma di grafo e coinvolti in esperimenti pubblicati in letteratura. Svolge il ruolo di libreria di supporto per la realizzazione di esempi dimostrativi e manuali inerenti alle altre librerie pubblicate. Disponibile nel repository pubblico di pacchetti R CRAN all'indirizzo:
  `http://cran.r-project.org/web/packages/bionetdata/index.html`.

E' pianificato nel breve periodo, compatibilmente con il ciclo di pubblicazone dei metodi, il rilascio di una libreria contenente i metodi sviluppati per l'ordinamento e la classificazione di nodi in grafi ed utilizzata per la realizzazione degli esperimenti pubblicati in[25, 7, 5].

II.F: Pubblicazione di dataset

Nell'ottica di promuovere il confronto di altri gruppi di ricerca con i risultati pubblicati è stata presa la decisione di pubblicare preferenzialmente in riviste internazionali open access ma anche di rendere pubblici i dataset utilizzati nelle procedure di valutazione delle performance dei metodi pubblicati.
A questo scopo sono stati resi pubblici:

- Un package R (BioNetData, descritto nella sottosezione precedente), contenente i dataset utilizzati in [**?**, 3].
- Tutti i dataset utilizzati negli esperimenti pubblicati in[2]. Essi sono disponibili a seguente indirizzo : `http://gigadb.org/dataset/100090` (DOI: `http://dx.doi.org/doi:10.1186/2047-217X-3-5`). Il DOI associato non coincide con quello della pubblicazione ma viene fornito automaticmente ad ogni dataset inserito nella banca dati pibblica GigaDB.

Pubblicazioni:

Articoli attualmente sottomessi a riviste internazionali con peer-review

- An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Articolo sottomesso a Genome Biology (impact factor: 10.8, ISSN: 1474-760X) contenente i risultati della seconda edizione della challenge internazionale Critical Assessment of protein Function Annotation - CAFA2
- RANKS: a flexible tool for node label ranking and classification in biological networks. Articolo sottomesso a Bioinformatics (impact factor: 4.981, ISSN: 1367-4803)

Riviste internazionali con peer-review

[1] G. Valentini, A. Paccanaro, H. Caniza, A.E. Romero, and **M. Re**. An estensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artificial Intelligence in Medicine*, 61:63–78, 2014. ISSN: 0933-3657. Impact factor: 2.019.

[2] M. Mesiti, **M. Re**, and G. Valentini. Think globally and solve locally: secondary memory-based network learning for automated multi-species functon prediction. *Gigascience*, 3:1–14, 2014. ISSN: 2047-217X.

[3] **M. Re** and G. Valentini. Network-based drug ranking and repositioning with respect to drugbank therapeutic categories. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10:1359–1371, 2013. ISSN: 1545-5963. Impact factor: 1.438.

[4] M. Frasca, A. Bertoni, **M. Re**, and G. Valentini. A neural network algorithm for semi-supervised node label learning from unbalanced data. *Neural Networks*, 43:84–98, 2013. ISSN: 0893-6080. Impact factor: 2.708.

[5] **M. Re** and G. Valentini. Cancer module genes ranking using kernelized score functions. *BMC Bioinformatics*, 13:S3–S3, 2012. ISSN: 1471-2105. Impact factor: 2.580.

[6] **M. Re**, M. Mesiti, and G. Valentini. A fast ranking algorithm for predicting gene functions in biomolecular networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9:1812–1818, 2012. ISSN: 1545-5963. Impact factor: 1.438.

[7] **M. Re**, M. Mesiti, and G. Valentini. Drug repositioning through pharmacological spaces integration based on networks projection. *EMBnet.journal*, 18:30–31, 2012. ISSN:2226-6089.

[8] A. Beghini, F. Corlazzoli, L. Del Giacco, **M. Re**, F. Lazzaroni, M. Brioschi, G. Valentini, F. Ferrazzi, A. Ghilardi, , M. Righi, M. Turrini, M. Mignardi, C. Cesana, V. Bronte, M. Nilsson, E. Morra, and R. Cairoli. Regeneration-associated wnt signaling is activated in long-term reconstituting ac133bright acute myeloid leukemia cells. *Neoplasia*, 14:1236–1248, 2012. ISSN: 1522-8002. Impact factor: 4.252.

[9] N. Cesa-Bianchi, **M. Re**, and G. Valentini. Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. *Machine Learning*, pages 1–33, December 2011. http://dx.doi.org/10.1007/s10994-011-5271-6 ISSN: 0885-6125. Impact factor: 1.889.

[10] **M. Ré** and G. Valentini. Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction. *Journal of Machine Learning Research - Machine Learning in Systems Biology*, 8:98–111, 2010. ISSN: 1938-7228.

[11] **M. Ré** and G. Valentini. Noise tolerance of multiple classifier systems in data integration-based gene function prediction. *Journal of Integrative Bioinformatics*, 7(3), 2010. ISSN: 1613-4516.

[12] **M. Re** and G. Valentini. Integration of heterogeneous data sources for gene function prediction using decision templates and ensembles of learning machines. *Neurocomputing*, 73(7-9):1533–1537, 2010. ISSN: 0925-2312. Impact factor: 2.083.

[13] **M. Ré**. Comparing early and late data fusion methods for gene expression prediction. *Soft Comput.*, 15(8):1497–1504, March 2010. ISSN: 1432-7643. Impact factor: 1.271.

[14] **M. Re**, G. Pesole, and D.S. Horner. Accurate discrimination of conserved coding and non-coding regions through multiple indicators of evolutionary dynamics. *BMC Bioinformatics*, 10:282, 2009. ISSN: 1471-2105. Impact factor: 2.580.

[15] **M. Ré** and G. Pavesi. Detecting conserved coding genomic regions through signal processing of nucleotide substitution patterns. *Artificial Intelligence in Medicine*, 45(2-3):117–123, 2009. ISSN: 0933-3657. Impact factor: 2.019.

[16] R. Avogadri, M. Brioschi, F. Ferrazzi, **M. Re**, A. Beghini, and G. Valentini. A stability-based algorithm to validate hierarchical clusters of genes. *IJKESDP*, 1(4):318–330, 2009.

[17] P. D'Onorio De Meo, D. Carrabino, N. Sanna, T. Castrignano, G. Grillo, F. Licciulli, S. Liuni, **M. Re**, F. Mignone, and G. Pesole. A high performance grid-web service framework for the identification of 'conserved sequence tags'. *Future Generation Comp. Syst.*, 23(3):371–381, 2007. ISSN: 0167-739X. Impact factor: 2.786.

[18] **M. Ré**, F. Mignone, M. Iacono, G. Grillo, S. Liuni, and G. Pesole. A new strategy to identify novel genes and gene isoforms: Analysis of human chromosomes 15, 21 and 22. *Gene*, (365):35–40, 2006. ISSN: 0378-1119. Impact factor: 2.138.

Editing di libri

[19] O. Okun, G. Valentini, and **M. Re**, editors. *Ensembles in Machine Learning Applications*, volume 373 of *Studies in Computational Intelligence*. Springer-Verlag Berlin Heidelberg, 2011. ISBN: 978-3-642-22909-1.

[20] O. Okun, **M. Re**, and G. Valentini, editors. *Ensembles in Machine Learning Applications*. Proceedings of the the Third Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications (SUEMA), European Conference on Machine Learning, Barcelona, Spain. 2010. Available as `http://suema10.dsi.unimi.it/suemafiles/SUEMA10_proceedings.pdf`.

Articoli di review in libri o riviste internazionali con peer review

[21] **M. Re** and G. Valentini. Ensemble methods: A review. In M.J Way, J.D. Scargle, K.M. Ali, and A.N. Srivastava, editors, *Advances in Machine Learning and Data Mining for Astronomy*, Data Mining and Knowledge Discovery, pages 563–582. Chapman and Hall an imprint of CRC Press (a division of Taylor and Francis), 2012. ISBN: 9781439841730.

Atti di conferenze internazionali e capitoli di libri con peer-review

[22] G. Valentini, S. Köhler, **M. Re**, M. Notaro, and P.N. Robinson. Prediction of human gene-phenotype associations by exploiting the hierarchical structure of the human phenotype ontology. In *Bioinformatics and biomedical engineering : third international conference, IWBBIO 2015, Granada, Spain, April 15-17,*

2015: proceedings., volume 9043 of *Lecture Notes in Computer Science*, pages 66–77. Springer Berlin / Heidelberg, 2015. ISBN: 978-3-319-16482-3.

[23] P.N. Robinson, M. Frasca, S. Köhler, M. Notaro, **M. Re**, and G. Valentini. A hierarchical ensemble method for dag-structured taxonomies. In *Multiple Classifier Systems : 12th international workshop MCS 2015, Günzburg, Germany, june 29 - july 1, 2015. Proceedings.*, volume 9132 of *Lecture Notes in Computer Science*, pages 15–26. Springer Berlin, 2015. ISBN: 978-3-319-20247-1.

[24] **M. Re** and G. Valentini. Large scale ranking and repositioning of drugs with respect to drugbank therapeutic categories. In Leonidas Bleris, Ion Mandoiu, Russell Schwartz, and Jianxin Wang, editors, *Bioinformatics Research and Applications*, volume 7292 of *Lecture Notes in Computer Science*, pages 225–236. Springer Berlin / Heidelberg, 2012.

[25] **M. Re** and G. Valentini. Genes prioritization with respect to cancer gene modules using functional linkage network data. In R. Bellazzi and P. Romano, editors, *11th International Workshop, NETTAB 2011, Network Tools and Application in Biology, Pavia, Italy, October 12-14, 2011, Proceedings*, pages 124–125, 2011.

[26] A. Rozza, G. Lombardi, **M. Re**, E. Casiraghi, G. Valentini, and P. Campadelli. A novel ensemble technique for protein subcellular location prediction. In O. Okun, G. Valentini, and M. Re, editors, *Ensembles in Machine Learning Applications*, volume 373 of *Studies in Computational Intelligence*, pages 151–167. Springer-Verlag Berlin Heidelberg, 2011.

[27] A. Bertoni, **M. Re**, F. Sacca, and G. Valentini. Identification of promoter regions in genomic sequences by 1-dimensional constraint clustering. In B. Apolloni, S. Bassis, A. Esposito, and C.F. Morabito, editors, *Neural Nets WIRN11 - Proceedings of the 21st Italian Workshop on Neural Nets, Vietri sul Mare, Salerno, Italy, 2011*, Frontiers in Artificial Intelligence and Applications, pages 162–169. IOS Press, 2011.

[28] **M. Re** and G. Valentini. An experimental comparison of hierarchical bayes and true path rule ensembles for protein function prediction. In N. El Gayar, J. Kittler, and F. Roli, editors, *Multiple Classifier Systems, 9th International Workshop, MCS 2010, Cairo, Egypt, April 7-9, 2010. Proceedings*, volume 5997 of *Lecture Notes in Computer Science*, pages 294–303, 2010.

[29] N. Cesa-Bianchi, **M. Re**, and G. Valentini. Functional inference in funcat through the combination of hierarchical ensembles with data fusion methods. In M. Zhang, G. Tsoumakas, and Z. Zhou, editors, *ICML/COLT Workshop on learning from Multi-Label Data MLD'10 Working Notes, Jun 25, Haifa, Israel*, pages 13–20, 2010. Available as `http://cse.seu.edu.cn/conf/mld10/files/MLD'10.pdf`.

[30] G. Valentini and **M. Re**. Weighted true path rule: a multilabel hierarchical algorithm for gene function prediction. In G. Tsoumakas, M. Zhang, and Z. Zhou, editors, *MLD-ECML 2009, 1st International Workshop on learning from Multi-Label Data, Sept 7, Bled, Slovenia*, pages 132–145, 2009. Available as `http://lpis.csd.auth.gr/workshops/mld09/mld09.pdf`.

[31] **M. Re** and G. Valentini. Simple ensemble methods are competitive with state-of-the-art data integration methods for gene function prediction. In S. Dzeroski, P. Geurts, and J. Rousu, editors, *Machine Learning in Systems Biology, Proceedings of the Third international workshop, Sept 5-6, Ljubljana, Slovenia*, pages 95–104, 2009.

[32] **M. Re** and G. Valentini. Prediction of gene function using ensembles of svms and heterogeneous data sources. In O. Okun and G. Valentini, editors, *Applications of Supervised and Unsupervised Ensemble Methods*, volume 245 of *Studies in Computational Intelligence*, pages 79–91. Springer, 2009.

[33] **M. Re** and G. Valentini. Predicting gene expression from heterogeneous data. In *Computational Intelligence Methods for Bioinformatics and Biostatistics - 6th International Meeting Proceedings, CIBB 2009, Genoa, Italy, October 15-17, 2009*, 2009.

[34] **M. Re** and G. Valentini. Ensemble based data fusion for gene function prediction. In J.A. Benediktsson, J. Kittler, and F. Roli, editors, *Multiple Classifier Systems, 8th International Workshop, MCS 2009, Reykjavik, Iceland, June 10-12, 2009. Proceedings*, volume 5519 of *Lecture Notes in Computer Science*, pages 448–457. Springer, 2009.

[35] **M. Re** and G. Valentini. Comparing early and late data fusion methods for gene function prediction. In B. Apolloni, S. Bassis, and F.C. Morabito, editors, *Neural Nets WIRN09 - Proceedings of the 19th Italian Workshop on Neural Nets, Vietri sul Mare, Salerno, Italy, May 28-30 2009*, Frontiers in Artificial Intelligence and Applications, pages 197–207. IOS Press, 2009.

[36] **M. Ré** and G. Pavesi. Signal processing in comparative genomics. In F. Masulli, S. Mitra, and G. Pasi, editors, *Applications of Fuzzy Sets Theory, 7th International Workshop on Fuzzy Logic and Applications, WILF 2007, Camogli, Italy, July 7-10, 2007, Proceedings*, volume 4578 of *Lecture Notes in Computer Science*, pages 544–550. Springer, 2007.

[37] **M. Re** and G. Valentini. Random walking on functional interaction networks to rank gene involved in cancer. In *Artificial Intelligence Applications and Innovations: AIAI 2012 International workshops: AIAB, AIeIA, CISE, COPA, IIVC, ISQL, MHDV and WADTMB, Halkidiki, Greece. September 27-30, 2012 : proceedings, Part II.*, volume 7292 of *Lecture Notes in Computer Science*, pages 225–236. Springer Berlin / Heidelberg, 2012.

Atti di conferenze nazionali

[38] A. Rozza, G. Lombardi, **M. Re**, E. Casiraghi, G. Valentini, and P. Campadelli. A novel ensemble approach for the subcellular localization of proteins. In *BITS 2011, Bioinformatics Italian Society Annual Meeting, Pisa, Italy, 2011. Proceedings*, 2011.

[39] D. Malchiodi, **M. Re**, and G. Valentini. Uso di mathematica per la classificazione di dati di qualit variabile. In *Mathematica Italia User Group Meeting - Atti del Convegno 2010*. Adalta, 2010.

[40] **M. Re** and G. Valentini. Data fusion based gene function prediction using ensemble methods. In *BITS 2009, Bioinformatics Italian Society Annual Meeting, Genova, Italy, 2009. Proceedings*, 2009.

[41] **M. Re**, C. Nasi, G. Pesole, and D.S. Horner. Efficient detection of conserved coding regions through a comparative genomic approach. In *BITS 2007, Bioinformatics Italian Society Annual Meeting, Napoli, Italy, 2007. Proceedings*, 2007.

[42] V. Piccolo, **M. Re**, G. Pesole, and S.D. Horner. Towards an integrated pipeline for the in-silico prediction of plant micrornas and their precursors. In *Nono Congresso annuale FISV (Federazione Italiana Scienze della Vita), Riva del Garda*, 2007.

[43] **M. Re**, S.D. Horner, C. Nasi, and G. Pesole. Improving the capacity of the cstminer algorithm to correctly classify conserved sequences. In *Ottavo Congresso annuale FISV (Federazione Italiana Scienze della Vita), Riva del Garda*, 2006.

[44] F. Mignone, **M. Re**, D.S. Horner, and G. Pesole. A new strategy to identify novel genes and genes isoforms: whole genome comparison of human and mouse. In *BITS 2006, Bioinformatics Italian Society Annual Meeting, Bologna, Italy, 2006. Proceedings*, 2006.

[45] D.S. Horner, **M. Re**, C. Nasi, and G. Pesole. Improving the cstminer algorithm to correctly classify conserved sequences. In *BITS 2006, Bioinformatics Italian Society Annual Meeting, Bologna, Italy, 2006. Proceedings*, 2006.

[46] **M. Re**, M. Iacono, F. Mignone, T. Castrignano, S. Liuni, G. Grillo, F. Licciulli, D.S. Horner, and G. Pesole. Identification of novel genes and genes isoforms in the human genome using cst miner, a novel algorithm for the differentiazion of coding and non-coding conserved sequence tags. In *BITS 2005, Bioinformatics Italian Society Annual Meeting, Milan, Italy, 2005. Proceedings*, 2005.